Towards Robust Linguistic Analysis Using OntoNotes

Sameer Pradhan[1], **Alessandro Moschitti**[2,3], **Nianwen Xue**[4], **Hwee Tou Ng**[5]
Anders Björkelund[6], **Olga Uryupina**[2], **Yuchen Zhang**[4] and **Zhi Zhong**[5]

[1] Boston Childrens Hospital and Harvard Medical School, Boston, MA 02115, USA
[2] University of Trento, University of Trento, 38123 Povo (TN), Italy
[3] QCRI, Qatar Foundation, 5825 Doha, Qatar
[4] Brandeis University, Brandeis University, Waltham, MA 02453, USA
[5] National University of Singapore, Singapore, 117417
[6] University of Stuttgart, 70174 Stuttgart, Germany

Abstract

Large-scale linguistically annotated corpora have played a crucial role in advancing the state of the art of key natural language technologies such as syntactic, semantic and discourse analyzers, and they serve as training data as well as evaluation benchmarks. Up till now, however, most of the evaluation has been done on monolithic corpora such as the Penn Treebank, the Proposition Bank. As a result, it is still unclear how the state-of-the-art analyzers perform in general on data from a variety of genres or domains. The completion of the OntoNotes corpus, a large-scale, multi-genre, multilingual corpus manually annotated with syntactic, semantic and discourse information, makes it possible to perform such an evaluation. This paper presents an analysis of the performance of publicly available, state-of-the-art tools on all layers and languages in the OntoNotes v5.0 corpus. This should set the benchmark for future development of various NLP components in syntax and semantics, and possibly encourage research towards an integrated system that makes use of the various layers jointly to improve overall performance.

1 Introduction

Roughly a million words of text from the Wall Street Journal newswire (WSJ), circa 1989, has had a significant impact on research in the language processing community — especially those in the area of syntax and (shallow) semantics, the reason for this being the seminal impact of the Penn Treebank project which first selected this text for annotation. Taking advantage of a solid syntactic foundation, later researchers who wanted to annotate semantic phenomena on a relatively large scale, also used it as the basis of their annotation. For example the Proposition Bank (Palmer et al., 2005), BBN Name Entity and Pronoun coreference corpus (Weischedel and Brunstein, 2005), the Penn Discourse Treebank (Prasad et al., 2008), and many other annotation projects, all annotate the same underlying body of text. It was also converted to dependency structures and other syntactic formalisms such as CCG (Hockenmaier and Steedman, 2002) and LTAG (Shen et al., 2008), thereby creating an even bigger impact through these additional syntactic resources. The most recent one of these efforts is the OntoNotes corpus (Weischedel et al., 2011). However, unlike the previous extensions of the Treebank, in addition to using roughly a third of the same WSJ subcorpus, OntoNotes also added several other genres, and covers two other languages — Chinese and Arabic: portions of the Chinese Treebank (Xue et al., 2005) and the Arabic Treebank (Maamouri and Bies, 2004) have been used to sample the genre of text that they represent.

One of the current hurdles in language processing is the problem of domain, or genre adaptation. Although genre or domain are popular terms, their definitions are still vague. In OntoNotes, "genre" means a type of source – newswire (NW), broadcast news (BN), broadcast conversation (BC), magazine (MZ), telephone conversation (TC), web data (WB) or pivot text (PT). Changes in the entity and event profiles across source types, and even in the same source over a time duration, as explicitly expressed by surface lexical forms, usually account for a lot of the decrease in performance of models trained on one source and tested on another, usually because these are the salient cues that are relied upon by statistical models.

Large-scale corpora annotated with multiple layers of linguistic information exist in various languages, but they typically consist of a single source or collection. The Brown corpus, which consists of multiple genres, have been usually used to investigate issues of genres of sensitivity, but it is relatively small and does not include any infor-

[1] A portion of the English data in the OntoNotes corpus is a selected set of sentences that were annotated for parse and word sense information. These sentences are present in a document of their own, and so the documents for parse layers for English are inflated by about 3655 documents and for the word sense are inflated by about 8797 documents.

Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 143–152,
Sofia, Bulgaria, August 8-9 2013. ©2013 Association for Computational Linguistics

Language	Parse		Proposition			Sense			Name		Coreference	
	Documents	Words	Documents	Verb Prop.	Noun Prop.	Documents	Verb Sense	Noun Sense	Documents	Words	Documents	Words
English	7,967[1]	2.6M	6,124	300K	18K	12K	173K	120K	3,637	2.0M	2,384 (3493)	1.7M
Chinese	2002	1.0M	1861	148K	7K	1573	83K	1K	1,911	988K	1,729 (2,280)	950K
Arabic	599	402K	599	30K	-	310	4.3K	8.7K	446	298K	447 (447)	300K

Table 1: Coverage for each layer in the OntoNotes v5.0 corpus, by number of documents, words, and some other attributes. The numbers in parenthesis are the total number of parts in the documents.

mal genres such as web data. Very seldom has it been the case that the exact same phenomena have been annotated on a broad cross-section of the same language before OntoNotes. The OntoNotes corpus thus provides an opportunity for studying the genre effect on different syntactic, semantic and discourse analyzers.

Parts of the OntoNotes Corpus have been used for various shared tasks organized by the language processing community. The word sense layer was the subject of prediction in two SemEval-2007 tasks, and the coreference layer was the subject of prediction in the SemEval-2010[2] (Recasens et al., 2010), CoNLL-2011 and 2012 shared tasks (Pradhan et al., 2011; Pradhan et al., 2012). The CoNLL-2012 shared task provided predicted information to the participants, however, that did not include a few layers such as the named entities for Chinese and Arabic, propositions for Arabic, and for better comparison of the English data with the CoNLL-2011 task, a smaller OntoNotes v4.0 portion of the English parse and propositions was used for training.

This paper is a first attempt at presenting a coherent high-level picture of the performance of various publicly available state-of-the-art tools on all the layers of OntoNotes in all three languages, so as to pave the way for further explorations in the area of syntax and semantics processing.

The possible avenues for exploratory studies on various fronts are enormous. However, given space considerations, in this paper, we will restrict our presentation of the performance on all layers of annotation in the data by using a stratified cross-section of the corpus for training, development, and testing. The paper is organized as follows: Section 2 gives an overview of the OntoNotes corpus. Section 3 explains the parameters of the evaluation and the various underlying assumptions. Section 4 presents the experimental results and discussion, and Section 5 concludes the paper.

2 OntoNotes Corpus

The OntoNotes project has created a large-scale corpus of accurate and integrated annotation of multiple layers of syntactic, semantic and discourse information in text. The English language portion comprises roughly 1.7M words and Chinese language portion comprises roughly 1M words of newswire, magazine articles, broadcast news, broadcast conversations, web data and conversational speech data[3]. The Arabic portion is smaller, comprising 300K words of newswire articles. This rich, integrated annotation covering many layers aims at facilitating the development of richer, cross-layer models and enabling better automatic semantic analysis. The corpus is tagged with syntactic trees, propositions for most verb and some noun instances, partial verb and noun word senses, coreference, and named entities. Table 1 gives an overview of the number of documents that have been annotated in the entire OntoNotes corpus.

2.1 Layers of Annotation

This section provides a very concise overview of the various layers of annotations in OntoNotes. For a more detailed description, the reader is referred to (Weischedel et al., 2011) and the documentation accompanying the v5.0[4] release.

2.1.1 Syntax

This represents the layer of syntactic annotation based on revised guidelines for the Penn Treebank (Marcus et al., 1993; Babko-Malaya et al., 2006), the Chinese Treebank (Xue et al., 2005) and the Arabic Treebank (Maamouri and Bies, 2004). There were two updates made to the parse trees as part of the OntoNotes project: i) the introduction of NML phrases, in the English portion, to mark nominal sub-constituents of flat NPs that do not follow the default right-branching structure, and ii) re-tokenization of hyphenated tokens into multiple tokens in English and Chinese. The Arabic Treebank on the other hand was also significantly revised in an effort to increase consistency.

2.1.2 Word Sense

Coarse-grained word senses are tagged for the most frequent polysemous verbs and nouns, in or-

[2] A small portion 125K words in English was used for this evaluation.

[3] These numbers are for the portion that has all layers of annotations. The word count for each layer is mentioned in Table 1

[4] For all the layers of data used in this study, the OntoNotes v4.99 pre-release that was used for the CoNLL-2012 shared task is identical to the v5.0 release.

der to maximize token coverage. The word sense granularity is tailored to achieve very high inter-annotator agreement as demonstrated by Palmer et al. (2007). These senses are defined in the sense inventory files. In the case of English and Arabic languages, the sense-inventories (and frame files) are defined separately for each part of speech that is realized by the lemma in the text. For Chinese, however the sense inventories (and frame files) are defined per lemma – independent of the part of speech realized in the text.

2.1.3 Proposition

The propositions in OntoNotes are PropBank-style semantic roles for English, Chinese and Arabic. Most English verbs and few nouns were annotated using the revised guidelines for the English PropBank (Babko-Malaya et al., 2006) as part of the OntoNotes effort. Some enhancements were made to the English PropBank and Treebank to make them synchronize better with each other: one of the outcomes of this effort was that two types of LINKs that represent pragmatic coreference (LINK-PCR) and selectional preferences (LINK-SLC) were added to the original PropBank (Palmer et al., 2005). More details can be found in the addendum to the PropBank guidelines[5] in the OntoNotes v5.0 release. A part of speech agnostic Chinese PropBank (Xue and Palmer, 2009) guidelines were used to annotate most frequent lemmas in Chinese. Many verbs and some nouns and adjectives were annotated using the revised Arabic PropBank guidelines (Palmer et al., 2008; Zaghouani et al., 2010).

2.1.4 Named Entities

The corpus was tagged with a set of 18 well-defined proper named entity types that have been tested extensively for inter-annotator agreement by Weischedel and Burnstein (2005).

2.1.5 Coreference

This layer captures general anaphoric coreference that covers entities and events not limited to noun phrases or a limited set of entity types (Pradhan et al., 2007). It considers all pronouns (PRP, PRP$), noun phrases (NP) and heads of verb phrases (VP) as potential mentions. Unlike English, Chinese and Arabic have dropped subjects and objects which were also considered during coreference annotation[6]. The mentions formed by these dropped pronouns total roughly about 11% for both Chinese and Arabic. Coreference is the only document-level phenomenon in OntoNotes. Some of the documents in the corpus — especially the ones in the broadcast conversation, web data, and telephone conversation genre — are very long which prohibited efficient annotation in their entirety. These are split into smaller parts, and each part is considered a separate document for the sake of coreference evaluation.

3 Evaluation Setting

Given the scope of the corpus and the multitude of settings one can run evaluations, we had to restrict this study to a relatively focused subset. There has already been evidence of models trained on WSJ doing poorly on non-WSJ data on parses (Gildea, 2001; McClosky et al., 2006), semantic role labeling (Carreras and Màrquez, 2005; Pradhan et al., 2008), word sense (Escudero et al., 2000; **?**), and named entities. The phenomenon of coreference is somewhat of an outlier. The winning system in the CoNLL-2011 shared task was one that was completely rule-based and not directly trained on the OntoNotes corpus. Given this overwhelming evidence, we decided not to focus on potentially complex cross-genre evaluations. Instead, we decided on evaluating the performance on each layer of annotation using an appropriately selected, stratified training, development and test set, so as to facilitate future studies.

3.1 Training, Development and Test Partitions

In this section we will have a brief discussion on the logic behind the partitioning of the data into training, development and test sets. Before we do that, it would help to know that given the range and peculiarities of the layers of annotation and presence of various resource and technical constraints, not all the documents in the corpus are annotated with all the layers of information, and token-centric phenomena (such as word sense and propositions of predicates) were not annotated with 100% coverage. Most of the proposition annotation in English and Arabic is for the verb predicates, with a few nouns annotated in English and some adjectives in Arabic. In Chinese, the selection is part of speech agnostic, and is based on the lemmas that can be considered predicates. Some documents in the corpora are actually snippets from larger documents, and have been annotated for a combination of parse, propositions, word sense and names, but not coreference. If one considers each layer independently, then an ideal partitioning scheme would create a separate partition for each layer such that it maximizes the number of examples that can be extracted for that layer from the corpus. The upside is that one would get as much data there is to train and estimate the performance of each layer across the entire corpus. The downside is that this might cover vari-

[5] doc/propbank/english-propbank.pdf

[6] As we will see later these are not used during the task.

ous cross sections of the documents in the corpus, and would not provide a clean picture when looking at the collective performance for all the layers. The documents that are annotated with coreference correspond to the intersection of all annotations. These are the documents that have also been annotated with all the other layers of information. The amount of data we can get together in such a test set is big enough to be representative. Therefore, we decided that it would be ideal to choose a portion of these documents as the test collection for all layers. An additional advantage is that it is the exact same test set used in the CoNLL-2012 shared task, and so in a way is already a standard. On the training and development side however, one can still imagine using all possible information for training models for a particular layer, and that is what we decided to do. The training and development data is generated by providing all documents with all available layers of annotation for input, however, the test set is generated by providing as input to the algorithm the set of documents in the corpus that have been annotated for coreference. This algorithm tries to reuse previously established partitions for English, i.e., the WSJ portion. Unfortunately, in the case of Chinese and Arabic, either the historical partitions were not in the selection used for OntoNotes, or were partially overlapping with the ones created using this scheme, and/or had a very small portion of OntoNotes covered in the test set. Therefore, we decided to create a fresh partition for the Chinese and Arabic data. Note, however, that the these test sets also match the ones used in the CoNLL-2012 evaluation. The algorithm for selecting the training, development and test partitions is described on the CoNLL-2012 shared task webpage, along with the list of training, development, and test document IDs[7].

3.2 Assumptions

Next we had to decide on a set of assumptions to use while designing the experiments to measure the automatic prediction accuracy for each of the layers. Since some of these decisions affect more than one layer of annotation, we will describe these in this section instead of in the section where we discuss the experiment with a particular layer of annotation.

Word Segmentation The three languages that we are evaluating are from quite different language families. Arabic has a complex morphology, English has limited morphology, whereas Chinese has very little morphology. English word segmentation amounts to rule-based tokenization, and is close to perfect. In the case of Chinese and Arabic, although the tokenization/segmentation is not as good as English, the accuracies are in the high 90s. Given this we decided to use gold, Treebank segmentation for all languages. In the case of Chinese, the words themselves are lemmas, whereas in English they can be predicted with very high accuracy. For Arabic, by default written text is unvocalised, and lemmatization is a complex process which we considered out of the scope of this study, so we decided to use correct, gold standard lemmas, along with the correct vocalized version of the tokens.

Traces and Function Tags Treebank traces have hardly played a role in the mainstream parser and semantic role labeling evaluation. Function tags also have received similar treatment in the parsing community, and though they are important, there is also a significant information overlap between them and the proposition structure provided by the PropBank layer. Whereas in English, most traces represent syntactic phenomena such as movement and raising, in Chinese and Arabic, they can also represent dropped subjects/objects. These subset of traces directly affect the coreference layer, since, unlike English, traces in Chinese and Arabic (***pro*** and * respectively) are legitimate targets of mentions and are considered for coreference annotation in OntoNotes. Recovering traces in text is a hard problem, and the most recently reported numbers in literature for Chinese are around a F-score of 50 (Yang and Xue, 2010; Cai et al., 2011). For Arabic there have not been much studies on recovering these. A study by Gabbard (2010) shows that these can be recovered with an F-score of 55 with automatic parses and roughly 65 using gold parses. Considering the low level of prediction accuracy of these tokens, and their relative low frequency, we decided to consider predicting traces in trees out of the scope of this study. In other words, we removed the manually identified traces and function tags from the Treebanks across all three languages, in all the three – training, development and test partitions. This meant removing any and all dependent annotation in layers such as PropBank and Coreference. In the case of PropBank these are the argument bearing traces, whereas in coreference these are the mentions formed by these elided subjects/objects.

[7] http://conll.cemantix.org/2012/download/ids/
For each language there are two sub-directories — "all" contains more general lists which include documents that had at least one of the layers of annotation, and "coref" contains the lists that include documents that have coreference annotation. The former were used to generate training, development, test sets for layers other than coreference, and the latter was used to generate training/development/test sets for the coreference layer used in the CoNLL-2012 shared task.

Disfluencies One thing that needs to be dealt with in conversational data is the presence of disfluencies (restarts, etc.). In the English parses of the OntoNotes, disfluencies are marked using a special EDITED[8] phrase tag – as was the case for the Switchboard Treebank. Computing the accuracy of identifying disfluencies is also out of the scope of this study. Given the frequency of disfluencies and the performance with which one can identify them automatically,[9] a probable processing pipeline would filter them out before parsing. We decided to remove them using oracle information available in the English Treebank, and the coreference chains were remapped to trees without disfluencies. Owing to various technical constraints, we decided to retain the disfluencies in the Chinese data.

Spoken Genre Given the scope of this study, we make another significant assumption. For the spoken genres – BC, BN and TC – we use the manual transcriptions rather than the output of a speech recognizer, as would be the case in real world. The performance on various layers for these genres would therefore be artificially inflated, and should be taken into account while analyzing results. Not many studies have previously reported on syntactic and semantic analysis for spoken genre. Favre et al. (2010) report the performance on the English subset of an earlier version of OntoNotes.

Discourse The corpus contains information on the speaker for broadcast communication, conversation, telephone conversation and writer for the web data. This information provides an important clue for correctly linking anaphoric pronouns with the right antecedents. This information could be automatically deduced, but is also not within the scope of our study. Therefore, we decided to provide gold, instead of predicted, data both during training and testing. Table 2 lists the status of the layers.

4 Experiments

In this section, we will report on the experiments carried out using all available data in the training set for training models for a particular layer, and using the CoNLL-2012 test set as the test set.

Layer	English	Chinese	Arabic
Segmentation	●	●	●
Lemma	○	—	●
Parse	○	○	○[10]
Proposition	○	○	○
Predicate Frame	○	○	○
Word Sense	○	○	○
Name Entities	○	○	○
Coreference	○	○	○
Speaker	●	●	—
Number	○	×	×
Gender	○	×	×

Table 2: Status of layers used during prediction of other layers. A "●" indicates gold annotation, a "○" indicates predicted, a "×" indicates an absence of the predicted layer, and a "—" indicates that the layer is not applicable to the language.

The predicted annotation layers input to downstream models were automatically annotated by using NLP processors learned with n-cross fold validation on the training data. This way, the n chunks of training data are annotated avoiding dependencies with the data used for training the NLP processors.

4.1 Syntax

Predicted parse trees for English were produced using the Charniak parser[11] (Charniak and Johnson, 2005). Some additional tag types used in the OntoNotes trees were added to the parser's tagset, including the nominal (NML) tag, and the rules used to determine head words were extended correspondingly. Chinese and Arabic parses were generated using the Berkeley parser (Petrov and Klein, 2007). In the case of Arabic, the parsing community uses a mapping from rich Arabic part of speech tags to Penn-style part of speech tags. We used the mapping that is included with the Arabic Treebank. The predicted parses for the training portion of the data were generated using 10-fold (5-folds for Arabic) cross-validation. For testing, we used a model trained on the entire training portion. Table 3 shows the precision, recall and F$_1$-scores of the re-trained parsers on the CoNLL-2012 test along with the part of speech accuracies (POS) using the standard `evalb` scorer.

The performance on the PT genre for English is the highest among other English genres. This is possibly because of the professional, clean translations of the underlying text, and are mostly shorter sentences. The MZ genre and the NW both of which contain well edited text, share similar scores. There is a few points gap between these and the other genres. As for Chinese, the performance on MZ is the highest followed by BN. Surprisingly, the WB genre has a similar score and the others are close behind except for TC. As expected, the Arabic parser performance is the low-

[8] There is another phrase type – EMBED in the telephone conversation genre which is similar to the EDITED phrase type, and sometimes identifies insertions, but sometimes contains logical continuation of phrases by different speakers, so we decided not to remove that from the data.

[9] A study by Charniak and Johnson (2001) shows that one can identify and remove edits from transcribed conversational speech with an F-score of about 78, with roughly 95 precision and 67 recall.

[10] The predicted part of speech for Arabic are a mapped down version of the richer gold version present in the Treebank

[11] http://bllip.cs.brown.edu/download/reranking-parserAug06.tar.gz

		All Sentences				
		N	POS	P	R	F
English	BC	2,211	97.33	86.36	86.11	86.23
	BN	1,357	97.32	87.61	87.03	87.32
	MZ	780	96.58	89.90	89.49	89.70
	NW	2,327	97.15	87.68	87.25	87.47
	TC	1,366	96.11	85.09	84.13	84.60
	WB	1,787	96.03	85.46	85.26	85.36
	PT	1,869	98.77	95.29	94.66	94.98
	Overall	11,697	97.09	88.08	87.65	87.87
Chinese	BC	885	94.79	80.17	79.35	79.76
	BN	929	93.85	83.49	80.13	81.78
	MZ	451	97.06	88.48	83.85	86.10
	NW	481	94.07	82.26	77.28	79.69
	TC	968	92.22	71.90	69.19	70.52
	WB	758	92.37	82.57	78.92	80.70
	Overall	4,472	94.12	82.23	78.93	80.55
Arabic	NW	1,003	94.12	74.71	75.67	75.19

Table 3: Parser performance on the CoNLL-2012 test set.

		Performance			
		P	R	F	A
English	BC	81.2	81.3	81.2	-
	BN	82.0	81.5	81.7	-
	MZ	79.1	78.8	79.0	-
	NW	85.7	85.7	85.7	-
	WB	77.5	77.6	77.5	-
	Overall	82.5	82.5	82.5	-
	Nouns	83.4	83.1	83.2	-
	Verbs	81.8	81.9	81.8	-
Chinese	BC	-	-	-	80.5
	BN	-	-	-	85.4
	MZ	-	-	-	82.4
	NW	-	-	-	89.1
	Overall	-	-	-	84.3
Arabic	NW	75.9	75.2	75.6	-
	Nouns	79.2	77.7	78.4	-
	Verbs	68.8	69.5	69.1	-

Table 4: Word sense performance on the CoNLL-2012 test set.

est among the three languages.

4.2 Word Sense

We used the IMS[12] (It Makes Sense) (Zhong and Ng, 2010) word sense tagger. IMS was trained on all the word sense data that is present in the training portion of the OntoNotes corpus using cross-validated predictions on the input layers similar to the proposition tagger. During testing, for English and Arabic, IMS must first use the automatic POS information to identify the nouns and verbs in the test data, and then assign senses to the automatically identified nouns and verbs. In the case of Arabic, IMS uses gold lemmas. Since automatic POS tagging is not perfect, IMS does not always output a sense to all word tokens that need to be sense tagged due to wrongly predicted POS tags. As such, recall is not the same as precision on the English and Arabic test data. For Chinese the measure of performance is just the accuracy since the senses are defined per lemma rather than per part of speech. Since we provide gold word segmentation, IMS attempts to sense tag all correctly segmented Chinese words, so recall and precision are the same and so is the F_1-score. Table 4 shows the performance of this classifier aggregated over both the verbs and nouns in the CoNLL-2012 test set and an overall score split by nouns and verbs for English and Arabic. For both nouns and verbs in English, the F_1-score is over 80%. The performance on English nouns is slightly higher than English verbs. Comparing to the other two languages, the performance on Arabic is relatively lower, especially the performance on Arabic verbs, whose F_1-score is less than 70%. For English, genres PT and TC, and for Chinese genres TC and WB, no gold standard senses were available, and so their accuracies could not be computed. Previously, Zhong et al. (2008) reported the word sense performance on the Wall Street Journal portion of an earlier version of OntoNotes, but the results are not directly comparable.

4.3 Proposition

The revised PropBank has introduced two new links — LINK-SLC and LINK-PCR. Since the community is not used to the new PropBank representation which (i) relies heavily on the trace structure in the Treebank and (ii) we decided to exclude, we *unfold* the LINKs back to their original representation as in the PropBank 1.0 release. We used ASSERT[15] (Pradhan et al., 2005) to predict the propositional structure for English. We made a small modification to ASSERT, and replaced the TinySVM classifier with a CRF[16] to speed up training the model on all the data. The Chinese propositional structure was predicted with the Chinese semantic role labeler described in (Xue, 2008), retrained on the OntoNotes v5.0 data. The Arabic propositional structure was predicted using the system described in Diab et al. (2008). (Diab et al., 2008) Table 5 shows the detailed per-

		Frame ID	Total Sent.	Total Prop.	% Perfect Prop.	Argument ID + Class		
						P	R	F
English	BC	93.2	1994	5806	52.89	80.76	69.69	74.82
	BN	92.7	1218	4166	54.78	80.22	69.36	74.40
	MZ	90.8	740	2655	50.77	79.13	67.78	73.02
	NW	92.8	2122	6930	46.45	79.80	66.80	72.72
	TC	91.8	837	1718	49.94	79.85	72.35	75.91
	WB	90.7	1139	2751	42.86	80.51	69.06	74.35
	PT	96.6	1208	2849	67.53	89.35	84.43	86.82
	Overall	92.8	9,261	26,882	51.66	81.30	70.53	75.53
Chinese	BC	87.7	885	2,323	31.34	53.92	68.60	60.38
	BN	93.3	929	4,419	35.44	64.34	66.05	65.18
	MZ	92.3	451	2,620	31.68	65.04	65.40	65.22
	NW	96.6	481	2,210	27.33	69.28	55.74	61.78
	TC	82.2	968	1,622	32.74	48.70	59.12	53.41
	WB	87.8	758	1,761	35.21	62.35	68.87	65.45
	Overall	90.9	4,472	14,955	32.62	61.26	64.48	62.83
Arabic	NW	85.6	1,003	2337	24.18	52.99	45.03	48.68

Table 5: Proposition and frameset disambiguation performance[14] in the CoNLL-2012 test set.

[12] http://www.comp.nus.edu.sg/∼nlp/sw/IMS_v0.9.2.1.tar.gz

[14] The Frame ID column indicates the F-score for English and Arabic, and accuracy for Chinese for the same reasons as word sense.

[15] http://cemantix.org/assert.html

[16] http://leon.bottou.org/projects/sgd

formance numbers[17]. The CoNLL-2005 scorer[18] was used to compute the scores. At first glance, the performance on the English newswire genre is much lower than what has been reported for WSJ Section 23. This could be attributed to several factors: i) the newswire in OntoNotes not only contains WSJ data, but also Xinhua news, and some other newswire evaluation data, ii) The WSJ training and test portions in OntoNotes are a subset of the standard ones that have been used to report performance earlier; iii) the PropBank guidelines were significantly revised during the OntoNotes project in order to synchronize well with the Treebank, and finally iv) it includes propositions for *be* verbs missing from the original PropBank. It looks like the newly added Pivot Text data (comprised of the New Testament) shows very good performance. The Chinese and Arabic[19] accuracy is much worse. In addition to automatically predicting the arguments, we also trained the IMS system to tag PropBank frameset IDs.

Language	Genre	Entity Count	Performance		
			P	R	F
English	BC	1671	80.17	77.20	78.66
	BN	2180	88.95	85.69	87.29
	MZ	1161	82.74	82.17	82.45
	NW	4679	86.79	84.25	85.50
	TC	362	74.09	61.60	67.27
	WB	1133	77.72	68.05	72.56
	Overall	11186	84.04	80.86	82.42
Chinese	BC	667	72.49	58.47	64.73
	BN	3158	82.17	71.50	76.46
	NW	1453	86.11	76.39	80.96
	MZ	1043	65.16	56.66	60.62
	TC	200	48.00	60.00	53.33
	WB	886	80.60	51.13	62.57
	Overall	7407	78.20	66.45	71.85
Arabic	NW	2550	74.53	62.55	68.02

Table 6: Performance of the named entity recognizer on the CoNLL-2012 test set.

4.4 Named Entities

We retrained the Stanford named entity recognizer[20] (Finkel et al., 2005) on the OntoNotes data. Table 6 shows the performance details for all the languages across all 18 name types broken down by genre. In English, BN has the highest performance followed by the NW genre. There is a significant drop from those and the TC and WB genre. Somewhat similar trend is observed in the Chinese data, with Arabic having the lowest scores. Since the Pivot Text portion (PT) of OntoNotes was not tagged with names, we could not compute the accuracy for that cross-section of the data. Previously Finkel and Manning (2009) performed a joint estimation of named entity and parsing. However, it was on an earlier version of the English portion of OntoNotes using a different cross-section for training and testing and therefore is not directly comparable.

4.5 Coreference

The task is to automatically identify mentions of entities and events in text and to link the coreferring mentions together to form entity/event chains. The coreference decisions are made using automatically predicted information on other structural and semantic layers including the parses, semantic roles, word senses, and named entities that were produced in the earlier sections. Each document part from the documents that were split into multiple parts during coreference annotation were treated as separate document.

We used the number and gender predictions generated by Bergsma and Lin (2006). Unfortunately neither Arabic, nor Chinese have comparable data available. Chinese, in particular, does not have number or gender inflections for nouns, but (Baran and Xue, 2011) look at a way to infer such information.

We trained the Björkelund and Farkas (2012) coreference system[21] which uses a combination of two pair-wise resolvers, the first is an incremental chain-based resolution algorithm (Björkelund and Farkas, 2012), and the second is a best-first resolver (Ng and Cardie, 2002). The two resolvers are combined by stacking, i.e., the output of the first resolver is used as features in the second one. The system uses a large feature set tailored for each language which, in addition to classic coreference features, includes both lexical and syntactic information.

Recently, it was discovered that there is possibly a bug in the official scorer used for the CoNLL 2011/2012 and the SemEval 2010 coreference tasks. This relates to the mis-implementation of the method proposed by (Cai and Strube, 2010) for scoring predicted mentions. This issue has also been recently reported in Recasens et al., (2013). As of this writing, the BCUBED metric has been fixed, and the correctness of the CEAF$_m$, CEAF$_e$ and BLANC metrics is being verified. We will be updating the CoNLL shared task webpages[22] with more detailed information and also release the patched scripts as soon as they are available. We will also re-generate the scores for previous shared tasks, and the coreference layer in this paper and make them available along with the models and system outputs for other layers. Table 7 shows the performance of the system on the

[17] The number of sentences in this table are a subset of the ones in the table showing parser performance, since these are the sentences for which at least one predicate has been tagged with its arguments

[18] http://www.lsi.upc.es/∼srlconll/srl-eval.pl

[19] The system could not not use the morphology features in Diab et al. (2008).

[20] http://nlp.stanford.edu/software/CRF-NER.shtml

[21] http://www.ims.uni-stuttgart.de/∼anders/coref.html

[22] http://conll.cemantix.org

CoNLL-2012 test set, broken down by genre. The same metrics that were used for the CoNLL-2012 shared task are computed, with the CoNLL column being the official CoNLL measure.

Language	Genre	MD	MUC	BCUBED	CEAF$_m$	CEAF$_e$	BLANC	CONLL
PREDICTED MENTIONS								
English	BC	73.43	63.92	61.98	54.82	42.68	73.04	56.19
	BN	73.49	63.92	65.85	58.93	48.14	72.74	59.30
	MZ	71.86	64.94	71.38	64.03	50.68	78.87	62.33
	NW	68.54	60.20	65.11	57.54	45.10	73.72	56.80
	PT	86.95	79.09	68.33	65.52	50.83	77.74	66.08
	TC	80.81	76.78	71.35	65.41	45.44	82.45	64.52
	WB	74.43	66.86	61.43	54.76	42.05	73.54	56.78
	Overall	75.38	67.58	65.78	59.20	45.87	75.8	59.74
Chinese	BC	68.02	59.6	59.44	53.12	40.77	73.63	53.27
	BN	68.57	61.34	67.83	60.90	48.10	77.39	59.09
	MZ	55.55	48.89	58.83	55.63	46.04	74.25	51.25
	NW	89.19	80.71	73.64	76.30	70.89	82.56	75.08
	TC	77.72	73.59	71.65	64.30	48.52	83.14	64.59
	WB	72.61	65.79	62.32	56.71	43.67	77.45	57.26
	Overall	66.37	58.61	66.56	59.01	48.19	76.07	57.79
Arabic	NW	60.55	47.82	61.16	53.42	44.30	69.63	51.09
GOLD MENTIONS								
English	BC	85.63	76.09	68.70	61.73	49.87	76.24	64.89
	BN	82.11	73.56	71.52	63.67	52.29	75.70	65.79
	MZ	85.65	77.73	78.82	72.75	60.09	83.88	72.21
	NW	80.68	73.52	73.08	65.63	51.96	81.06	66.19
	PT	93.20	85.72	73.25	70.76	58.81	79.78	72.59
	TC	90.68	86.83	78.94	73.87	56.26	85.82	74.01
	WB	88.12	80.61	69.86	63.45	51.13	76.48	67.20
	Overall	86.16	78.7	72.67	66.32	53.23	79.22	68.2
Chinese	BC	84.88	76.34	69.89	62.02	49.29	76.89	65.17
	BN	80.97	74.89	76.88	68.91	55.56	81.94	69.11
	MZ	78.85	73.06	70.15	61.68	46.86	78.78	63.36
	NW	93.23	86.54	86.70	80.60	76.60	85.75	83.28
	TC	92.91	88.31	84.51	79.49	63.87	90.04	78.90
	WB	85.87	77.61	69.24	60.71	47.47	77.67	64.77
	Overall	83.47	76.85	76.30	68.30	56.61	81.56	69.92
Arabic	NW	76.43	60.81	67.29	59.50	49.32	74.61	59.14

Table 7: Performance of the coreference system on the CoNLL-2012 test set.

The varying results across genres mostly meet our expectations. In English, the system does best on TC and the PT genres. The text in the TC set often involve long chains where the speakers refer to themselves which, given speaker information, is fairly easy to resolve. The PT section includes many references to god (e.g. *god* and *the lord*) which the lexicalized resolver is quite good at picking up during training. The more difficult genres consist of texts where references to many entities are interleaved in the discourse and is as such harder to resolve correctly. For Chinese the numbers on the TC genre are also quite good, and the explanation above also holds here — many mentions refer to either of the speakers. For Chinese the NW section displays by far the highest scores, however, and the reason for this is not clear to us. Not surprisingly, restricting the set of mentions only to gold mentions gives a large boost across all genres and all languages. This shows that mention detection (MD) and singleton detection (which is not part of the annotation) remain a big source of errors for the coreference resolver. For these experiments we used a combination of training and development data for training — following the CoNLL-2012 shared task specification. Leaving out the development set has a very negligible effect on the CoNLL-score for all the languages (English: 0.14; Chinese 0.06; Arabic: 0.40 F-score respectively). The effect on Arabic is the most (0.40 F-score) most likely because of its much smaller size. To gauge the performance improvement between 2011 and 2012 shared tasks, we performed a clean comparison of over the best performing system and an earlier version of this system (Björkelund and Nugues, 2011) on the CoNLL 2011 test set using the CoNLL 2011 train and development set for training. The current system has a CoNLL score of 60.09 ($\frac{64.92+69.84+45.51}{3}$)[23] as opposed to the 54.53 reported in *björkelund* (Björkelund and Nugues, 2011), and the 57.79 reported for the best performing system of CoNLL-2011. One caveat is that these score comparison are done using the earlier version (v4) of the CoNLL scorer. Nevertheless, it is encouraging to see that within a short span of a year, there has been significant improvement in system performance – partially owing to cross-pollination of research generated through the shared tasks.

5 Conclusion

In this paper we reported work on finding a reasonable training, development and test split for the various layers of annotation in the OntoNotes v5.0 corpus, which consists of multiple genres in three typologically very different languages. We also presented the performance of publicly available, state-of-the-art algorithms on all the different layers of the corpus for the different languages. The trained models as well as their output will be made publicly available[24] to serve as benchmarks for language processing community. Training so many different NLP components is very time-consuming, thus, we hope the work reported here has lifted the burden of having to create reasonable baselines for researchers who wish to use this corpus to evaluate their systems. We created just one data split in training, development and test set, covering a collection of genres for each layer of annotation in each language in order to keep the workload manageable However, the results do not discriminate the performance on individual genres: we believe such a setup is still a more realistic gauge for the performance of the state-of-the-art NLP components than a monolithic corpus such as the Wall Street Journal section of the Penn Treebank. It can be used as a starting point for developing the next generation of NLP components that are more robust and perform well on a multitude of genres for a variety of different languages.

[23] (MUC + BCUBED + CEAF$_e$)/3

[24] http://cemantix.org

6 Acknowledgments

We gratefully acknowledge the support of the Defense Advanced Research Projects Agency (DARPA/IPTO) under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022 for sponsoring the creation of the OntoNotes corpus. This work was partially supported by grants R01LM10090 and U54LM008748 from the National Library Of Medicine, and R01GM090187 from the National Institutes of General Medical Sciences. We are indebted to Slav Petrov for helping us to retrain his syntactic parser for Arabic. Alessandro Moschitti and Olga Uryupina have been partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under the grant number 288024 (LIMOSINE). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Nianwen Xue and Yuchen Zhang are supported in part by the DAPRA via contract HR0011-11-C-0145 entitled "Linguistic Resources for Multilingual Processing."

References

Olga Babko-Malaya, Ann Bies, Ann Taylor, Szuting Yi, Martha Palmer, Mitch Marcus, Seth Kulick, and Libin Shen. 2006. Issues in synchronizing the English treebank and propbank. In *Workshop on Frontiers in Linguistically Annotated Corpora 2006*, July.

Elizabeth Baran and Nianwen Xue. 2011. Singular or plural? exploiting parallel corpora for Chinese number prediction. In *Proceedings of Machine Translation Summit XIII*, Xiamen, China.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Sydney, Australia, July.

Anders Björkelund and Richárd Farkas. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 49–55, Jeju Island, Korea, July. Association for Computational Linguistics.

Anders Björkelund and Pierre Nugues. 2011. Exploring lexicalized features for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 45–50, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 28–36.

Shu Cai, David Chiang, and Yoav Goldberg. 2011. Language-independent parsing with empty elements. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 212–216, Portland, Oregon, USA, June. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL)*, Ann Arbor, MI, June.

Eugene Charniak and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, June.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, June.

Mona Diab, Alessandro Moschitti, and Daniele Pighin. 2008. Semantic role labeling systems for Arabic using kernel methods. In *Proceedings of ACL-08: HLT*, pages 798–806, Columbus, Ohio, June. Association for Computational Linguistics.

Gerard Escudero, Lluis Marquez, and German Rigau. 2000. An empirical study of the domain dependence of supervised word disambiguation systems. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 172–180, Hong Kong, China, October. Association for Computational Linguistics.

Benoit Favre, Bernd Bohnet, and D. Hakkani-Tur. 2010. Evaluation of semantic role labeling and dependency parsing of automatic speech recognition output. In *Proceedings of 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 5342–5345.

Jenny Rose Finkel and Christopher D. Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 326–334, Boulder, Colorado, June. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, page 363–370.

Ryan Gabbard. 2010. *Null Element Restoration*. Ph.D. thesis, University of Pennsylvania.

Daniel Gildea. 2001. Corpus variation and parser performance. In *2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Pittsburgh, PA.

Julia Hockenmaier and Mark Steedman. 2002. Acquiring compact lexicalized grammars from a cleaner treebank. In *Proceedings of the Third LREC Conference*, page 1974–1981.

Mohamed Maamouri and Ann Bies. 2004. Developing an Arabic treebank: Methods, guidelines, procedures, and tools. In Ali Farghaly and Karine Megerdoomian, editors, *COLING 2004 Computational Approaches to Arabic Script-based Languages*, pages 2–9, Geneva, Switzerland, August 28th. COLING.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, June.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, New York City, NY, June.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the Association for Computational Linguistics (ACL-02)*, pages 104–111.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*, 13(2).

Martha Palmer, Olga Babko-Malaya, Ann Bies, Mona Diab, Mohammed Maamouri, Aous Mansouri, and Wajdi Zaghouani. 2008. A pilot Arabic propbank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 28-30.

Slav Petrov and Dan Klein. 2007. Improved inferencing for unlexicalized parsing. In *Proc of HLT-NAACL*.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James Martin, and Dan Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Indentifying entities and events in OntoNotes. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*, September 17-19.

Sameer Pradhan, Wayne Ward, and James H. Martin. 2008. Towards robust semantic role labeling. *Computational Linguistics Special Issue on Semantic Role Labeling*, 34(2).

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, July. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden, July.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 627–633, Atlanta, Georgia, June. Association for Computational Linguistics.

Libin Shen, Lucas Champollion, and Aravind K. Joshi. 2008. LTAG-spinal and the treebank. *Language Resources and Evaluation*, 42(1):1–19, March.

Ralph Weischedel and Ada Brunstein. 2005. BBN pronoun coreference and entity type corpus LDC catalog no.: LDC2005T33. BBN Technologies.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143–172.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Nianwen Xue. 2008. Labeling Chinese predicates with semantic roles. *Computational Linguistics*, 34(2):225–255.

Yaqin Yang and Nianwen Xue. 2010. Chasing the ghost: recovering empty categories in the Chinese treebank. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China.

Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. 2010. The revised Arabic propbank. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 222–226, Uppsala, Sweden, July.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden.

Zhi Zhong, Hwee Tou Ng, and Yee Seng Chan. 2008. Word sense disambiguation using OntoNotes: An empirical study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1002–1010.